Research Data Management
Birkbeck, University of London

Open Research Survey 2017

Report

Title:	Open Research Survey 2017 - Report
Version:	1.2
Author:	David McElroy (https://orcid.org/0000-0002-0966-8862), Sarah Lee
Licence:	CC-BY
Update Date:	14-12-2017

Contents

Introduction

Providing Research Data Management (RDM) support has many challenges. The different types of data, the volume storied on various media, the differences across disciplines, all contribute to the complexity of supporting RDM in a diverse institution such as Birkbeck.

Another part of the challenge of delivering an RDM service is the varying attitudes toward data management. Staff are often unclear of what "Research Data" means to them, and have often have mixed levels of previous support or training.

Without knowing what our academics already understand of RDM, it would be very hard to meet their needs in the correct way.

Many institutions have run these types of surveys in the past few years, yielding useful results. Recently UCL, SOAS, and a consortium led by Jisc have run very similar surveys, which we hoped to be able to compare our results to.

Open Access was also included in the survey; so as to compare attitudes to a previous survey which ran in 2011. Same questions were deliberately the same, while we added some extra questions about the Research Excellence Frame work to gauge current staff attitudes.

Objectives

We considered the following questions at the beginning of this survey:

1. How have attitudes towards Open Access changed since 2011?
2. What are the current attitudes towards Open Data and RDM?
3. How aware are Birkbeck researchers regarding internal and external RDM policies?
4. What are Birkbeck researchers current RDM practices? (This includes details on volume, security, location, accessibility, etc..)
5. What are the training needs that our researchers think they have, and what needs can be uncovered?
6. How does Birkbeck compare with other institutions in this area? Other HEIs have conduced similar surveys allowing some comparisons to be made.

The other major objective or the survey was to raise awareness of RDM, and the RDM support service provided by the library.

Data Asset Framework

The Data Asset Framework (DAF) was originally developed by the DCC and the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow.

From their website, “The Data Asset Framework (DAF) provides organisations with the means to identify, locate, describe and assess how they are managing their research data assets.” It does this by providing a survey methodology, and an implementation guide.

Further information on the Data Asset Framework is available on their website: http://www.data-audit.eu/

Jisc recently published an updated set of questions for a DAF, along with the results of survey. We can use the results of their surveys at the institutions they partnered with to complete objective number 6, along with the results of a simpler survey published by UCL in 2016.

Methods

Survey tool

The survey was run using the Bristol Online Survey tool (onlinesurveys.ac.uk). The 2011 survey was also conducted using this tool allowing for easy cross interrogation of the data.

Distribution

After discussions at the various boards who approved this survey, it was decided that the best route for distribution was by the Director of Library Services sending out a set text to the School Managers.

The School Managers then distributed the text and the survey link to their staff.

The library also tweeted about the survey, as did the Birkbeck Research Data account. This was retweeted by the Schools and Departmental accounts.

Response Rate

By looking at the response rates from previous surveys at other larger institutions, we could make a prediction on what a good response rate would be.

Institution	Date	Responses	Percentage (%)
LSHTM	2013	117	16.25[1]
Exeter	2012	284[2]	22.6*
Sheffield	2014	432	8[3]
St George's	2015	86[4]	
Cambridge	2016	440[5]	6.6*

*estimate

As the table shows, response rates vary between above 5% to mid-20%. After looking at the number of questions asked, we could not conclusively determine if the total number of questions asked had any relation to final response rate.

On close of the Birkbeck Open Research Survey 2017, we had a total of 85 responses.

Respondent Type	Count
Early Career Researcher (PhD completed in the last 5 years)	5
Experienced Researcher (PhD completed more than 5 years ago)	57
Research Student	21
Other	2

From these we can see that we had roughly 15% response rate from our staff which represents a reasonable response rate when we look at the responses to the other previous surveys listed above.

For the majority of this report, we will be combining our staff and PhD student's responses. PhD students are an important part of the research environment at The College, and their attitudes and practices are important to us. For some questions, we will exclude their responses, focusing on the staff members only. Where this is the case we will say so in the comments or in the title of the question being discussed.

[1] (Knight, 2013)
[2] (Open Exeter Project Team, 2012)
[3] (Cox and Williamson, 2015)
[4] (Basford, 2016)
[5] (Johnson, Chiarelli and Parsons, 2016)

The Survey

Our survey was laid out in sections, with each section taking up a page. This allowed us to organise the questions in a way that made sense to the respondents.

For the full list of questions asked, please see Appendix A

Introduction

In the opening section, we stated who the survey was for, the reasons behind running it, and what we define research data as. We also include an agreement on anonymity and data sharing.

You and Research

Here we ask questions about the researcher. From knowing about their School and Department and research experience, we can combine the responses with later questions to examine wither subject areas and experience influence practices and opinions on RDM and OA.

Open Access

As a previous survey was conducted in 2011, we decided that running the same questions could reveal useful inside into changing attitudes towards OA.

Research Data Management

In this section, we ask very similar questions to the previous OA section, with the intention of comparing attitudes toward OA and RDM/Open Data.

We also ask about the very important topic of training. This should directly impact the workshops we plan to run at Birkbeck, or potentially as part of the Bloomsbury Group.

Current Research (Introduction) & Current Research (Describing your data)

We asked questions relating to the respondent's current research project(s) to gain a greater understanding of what sort of research is currently being undertaken at Birkbeck.

The questions look at the type, volume, scale, vulnerability, and security of data. This will provide a very valuable insight into RDM practices at The College.

Past Research

Questions in this section give us an idea of the existing data that is being stored in Birkbeck drives and devices. It allows us to estimate more accurately the volume of data that is currently at risk if not backed up, and what percentage of this data could be placed in the institutional data repository BiRD.

General RDM (Sharing, Collaborations, and Details)

Here we gather more general information on RDM practices, which we can combine with other questions in previous sections to draw conclusions on how practices vary with discipline.

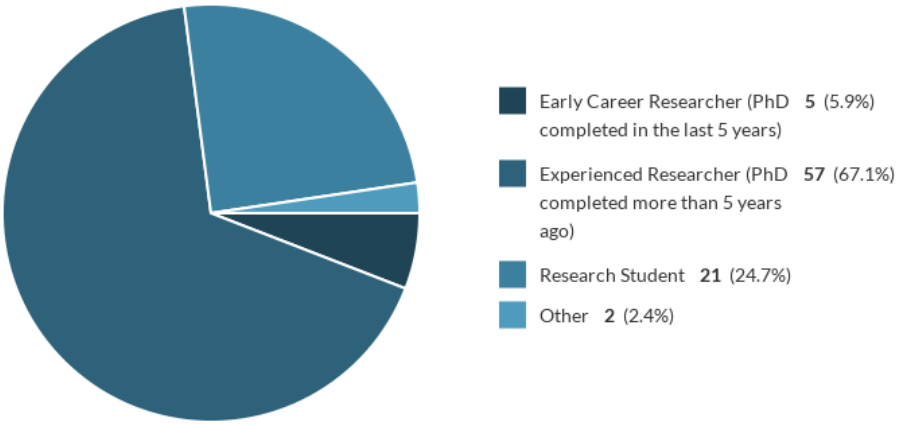
Final Questions

The final questions ask if there is any data the respondent would like to deposit, and if they have any other comments on RDM at The College.

Outcomes

Research Experience and School Engagement

Q3. Which of the following best describes your research experience?



From the figure above we can see that the majority of respondents to our survey were researchers with more than 5 years' experience. The second largest group being students. The "Other" reported were one research centre coordinator and one research assistant.

Looking at the breakdown of staff across the schools we see a good and even response from all Schools, with the exception of the School of Law.

School	Count
School of Arts	12
School of Business, Economics and Informatics	17
School of Law	1
School of Science	15
School of Social Sciences, History and Philosophy	19

For students, we see a similar spread of results, with Law again being weakest.

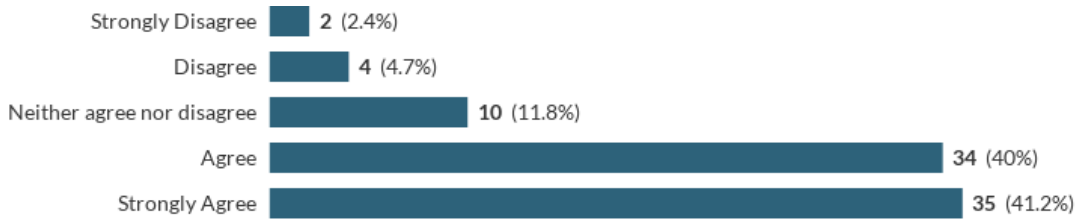
School	Count
School of Arts	4
School of Business, Economics and Informatics	2
School of Law	1
School of Science	8
School of Social Sciences, History and Pholosophy	6

Looking at the data from institute and centres, we see that 50 of the 85 responses came from researcher or students associated with at least one institute or centre. Some reported they were associated with more than one.

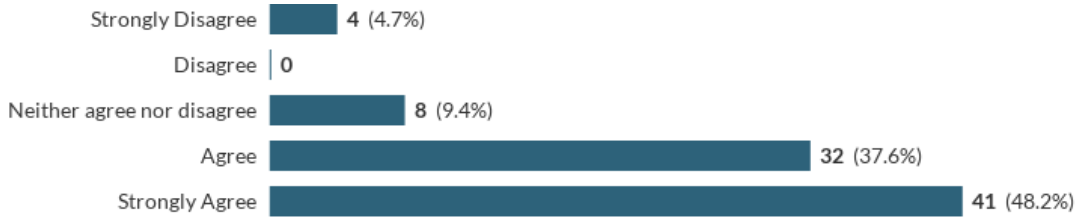
Open Access

As we stated previously, the decision to include the Open Access questions was made to allow us to compare the results with a previous OA study from 2011. The data from this survey is available in the Birkbeck Research Data repository (BiRD) here: https://doi.org/10.18743/data.00012

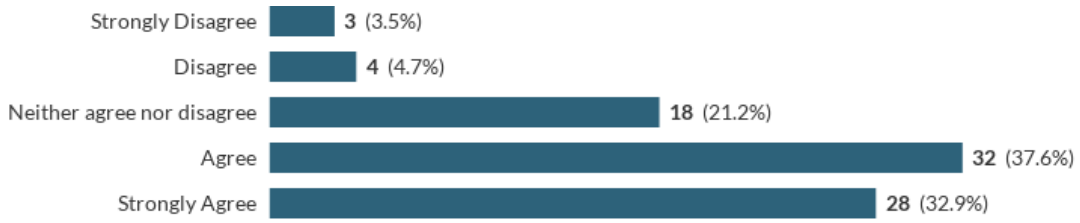
Q4.1 How do you feel about the principles of Open Access?



Q4.2 How do you feel about using Open Access repositories? (Repositories which make versions of articles freely available)



Q4.3 How do you feel about publishing in Open Access journals? (Journals which do not restrict access to articles)



We can then use the responses above, and compare them to the previous survey:

Question 4 (Staff)	2011	2017	Change	2011	2017	Change
	Agree or Strongly Agree			Disagree or Strongly Disagree		
How do you feel about the principles of Open Access?	89.4	81.3	-8.1	3	6.3	3.3
How do you feel about using Open Access repositories? (Repositories which make versions of articles freely available)	83.3	87.5	4.2	3	4.7	1.7
How do you feel about publishing in Open Access journals? (Journals which do no restrict access to articles)	63.1	68.7	5.6	15.4	7.8	-7.6

Question 7 & 8 (Staff)	2011	2017	Change	2011	2017	Change
	Yes			No/Not Sure		
Are you aware of The College repository, BIROn?	43.9	96.9	53	56.1	3.1	-53
Do you currently make any of your publications available in BIROn?	46.7	91.9	45.2	53.3	8	-45.3
Do you deposit your own publications? (self-archive)	12.5	79.7	67.2	87.5	20.3	-67.2

From the two tables above, Staff seem to be generally slightly less enthusiastic about the principles of Open Access.

This may be because in 2011:

1. Open Access was not a requirement in the same way it is for the current REF exercise
2. Hybrid Journals were not as common.

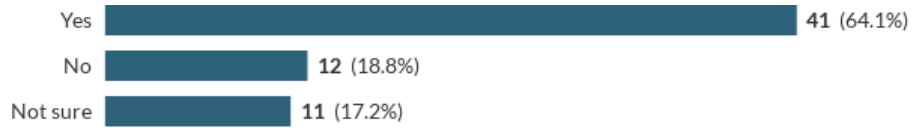
More positively, we see far more awareness of the repository BIROn. 53% more staff are now aware of its existence. Of those that have heard of it, 45% more deposit, with 67% depositing themselves.

This shows that the efforts of the Library to raise awareness, and increase use of the repository have been working. It also suggests that staff are aware the need to use the repository (for certain types of output) to be included in the next REF.

REF

Q5. Do you feel you understand the Open Access requirements for inclusion in the next REF?

Q6. Do you feel you understand the differences between Gold and Green Open Access?



With 17.2% who do not understand the requirement for the next REF, and the 36% (No + Not Sure) who don't feel they understand the difference between Gold and Green Open Access, more ongoing training and support is needed.

Data Attitudes

By asking similar questions to those described above, we can make comparisons between attitudes toward Open Access and Open Data.

Q9. Who do you think "should" own the copyright of research publications?

Q10. Who do you think "should" own the copyright of Research Data?

As we can see, there is a very slight difference between OA and Data when it comes to ownership.

"Other" for Q9 were split between the funders and journal publishers.

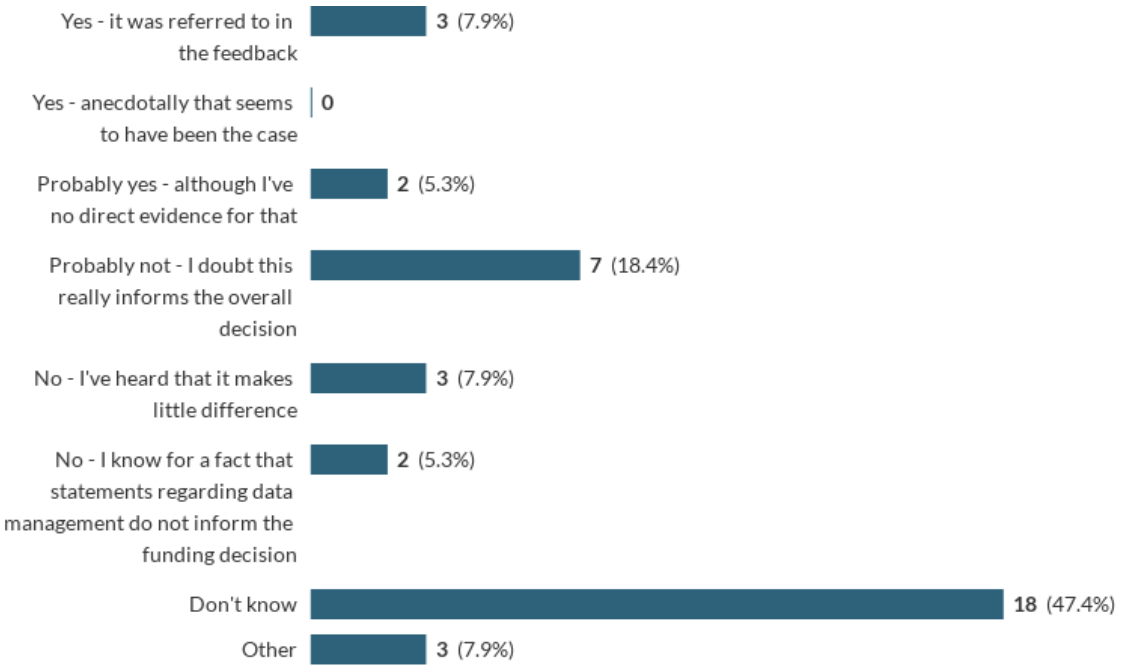
"Other" for Q10 were split between funders and don't knows.

It's hard to make any assumptions based on such similarities, other than to say that our staff expect the College to make similar provisions for the ownership of their data as they do for publication of results. These answers also highlight a common misconception about the ownership of copyright for research publications – where most authors currently sign over the copyright as part of the process of submitting to the publisher.

Data Management Planning is an important part of Research Data Management. There are many well documented benefits to creating a DMP[6]. However, there have been questions raised of how the funders review the need for these plans as part of funder applications, and whither they really value them as a project deliverable.

[6] (Jones, 2011)

Q15. If you have previously completed a DMP/Technical Plan/Data Sharing Plan, as part of an application for funding, do you think your funder seriously considered your responses before reaching a decision regarding the bid?



The majority of responses to this question were “Don’t know”, suggesting that there is a lot of uncertainty around DMPs, and how the funders see them. If we only include those who are funded by funders who are now asking for a DMP at some stage, we see similar percentages, which suggests that the funders too could do more to improve feedback from DMPs.

Finally, we look at whether our researcher think that data should be shared openly if possible.

Q11. Do you think Research Data should be shared openly where possible? (Data that is not commercial sensitive, or impossible to anonymise)

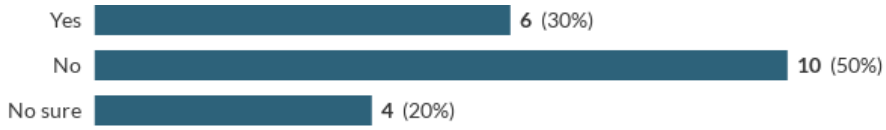
The strong result here shows that openly sharing data is accepted among the majority of Birkbeck researchers.

By splitting Question 11 into Early Career and PhD, and Senior Researchers, we see no real difference in response. The idea that younger researchers are more likely to share their data is not backed up by the results of our survey.

Policy Awareness

We have three ways to look at policy awareness in our survey. The first is to look at the research funders who mandate a DMP, and how many of those respondents have a DMP.

Q18. Do you have a data management plan (DMP) for your current research?

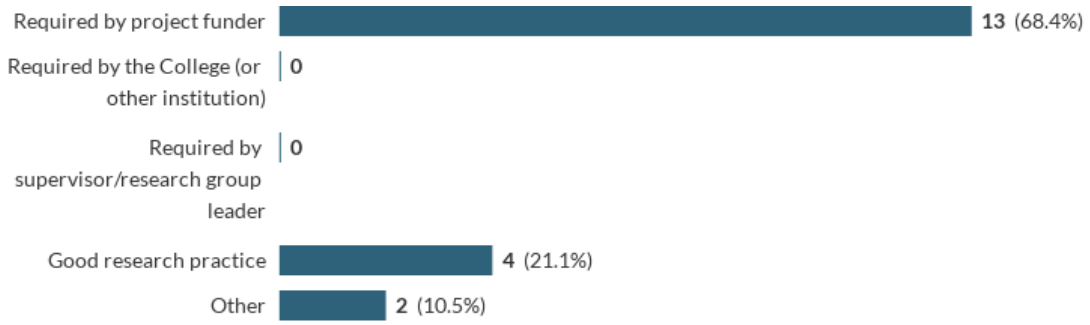


While some project maybe have started before the funder RDM policies and College policy were in place, most policies are a few years old now (EPSRC was endorsed in 2011 and mandated 2015, NERC 2014). The rules are not applied retrospectively, so many of our projects will be exempt from these requirements. We must still be cautious going forward as a significant number of current projects have no DMP.

Secondly, we can look at how many of our respondents mentioned they had created a DMP during their time at Birkbeck.

Q14. Birkbeck library offers support with Data Management Plans, for funded and non-funded research at The College. Have you ever created a Data Management Plan, Technical Plan, or Data Sharing Plan for any research you have undertaken here at Birkbeck?

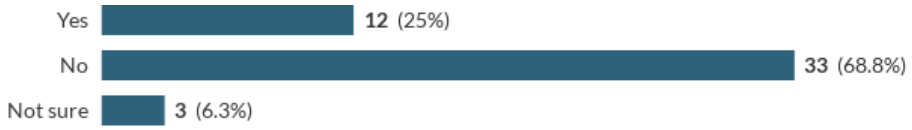
Q14b. What are your reasons for having created a Data Management Plan?



From the answer above, we see that not one respondent said they had created a DMP due to The College requiring one, despite a policy of mandating DMP for all funded research since 2016.

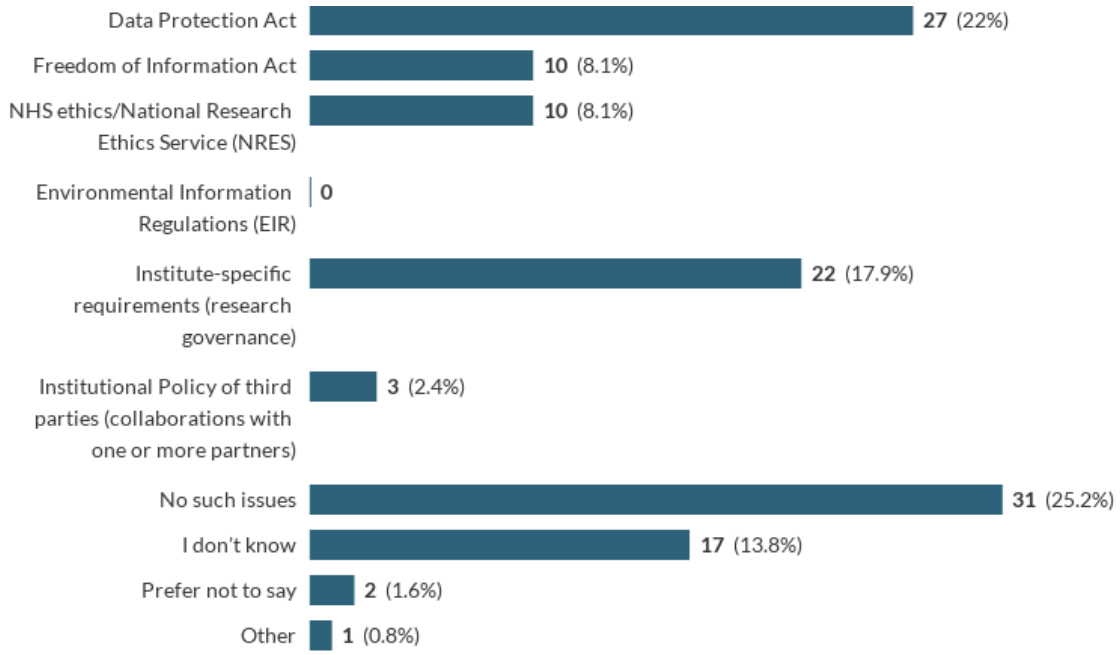
By limited the responses to those who are currently undertaking funded research, we can see that nearly 70% still do not have a DMP.

Q14 As above (limited to funded currently research)



Finally, we explicitly ask if they are aware of any policy that influences their research.

Q24. What, if any, legislation policies or other rules influence how your Research Data is stored, managed and/or shared?



From this question, we can see that just under 18% though there were College policy that might impact their data activities. This shows a very low rate of awareness for The College RDM policy.

This suggests that the RDM policy is not being communicated well enough at present and there is a widespread lack of understanding across the research base.

With further changes to policy coming in the shape of the GDPR[7], and further implementation of the Concordat on Open Research Data[8], we need to ensure that we have simple and clear guidance to

[7] https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/
[8] http://www.rcuk.ac.uk/media/news/160728/

allow our academics to get on with their research, while not being bogged down in compliance issues.

Current Data Practices

We wanted to gain an insight into how our researcher are currently storing their data.

Initially we asked about the nature of the data being created.

- **62.4%** stated their data was primary data
- **55.3%** stated their data contained quantitative data
- **55.3%** stated their data contained qualitative data

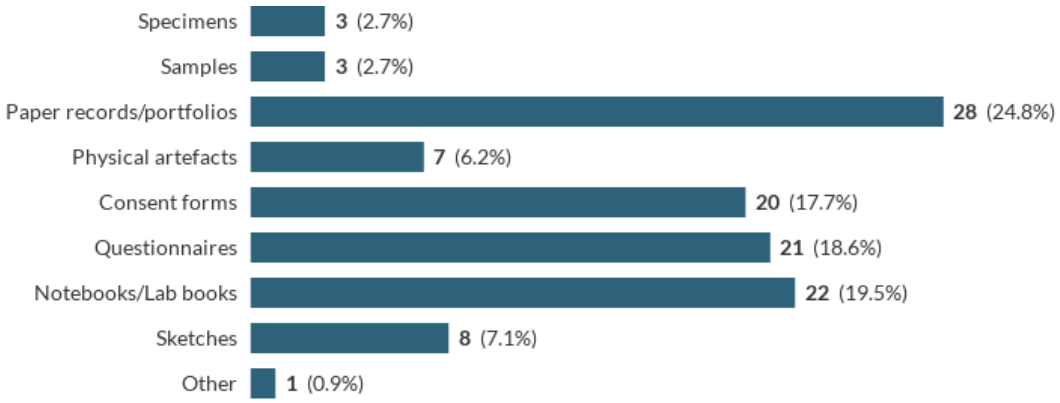
30 of the respondents said they worked with both quantitative and qualitative data.

Non-Digital

Over half of our researchers store non-digital data. This is common across discipline, with a variety of possibilities selected:

Q20. Do you store any non-digital Research Data? (e.g. notebooks, physical samples, field notes, etc..)

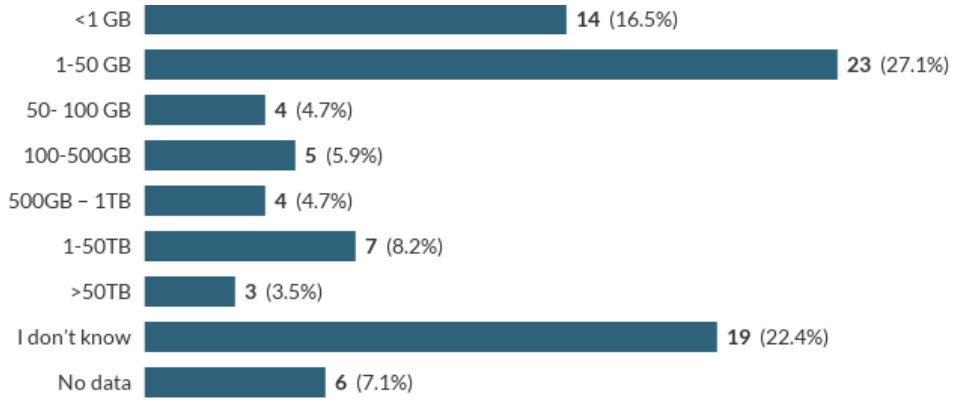
Q20a. What kinds of non-digital Research Data do you store? (select all that apply)



We follow this by asking if digital copies were created of this data, only 1/3 saying they did this. This suggests that around 30% of all data created at Birkbeck is no in a state to be easily shared digitally, regardless of whether the data can be shared for ethical/sensitivity reasons.

Data Volume and Scale

Q21. Estimate what volume of data are you creating?

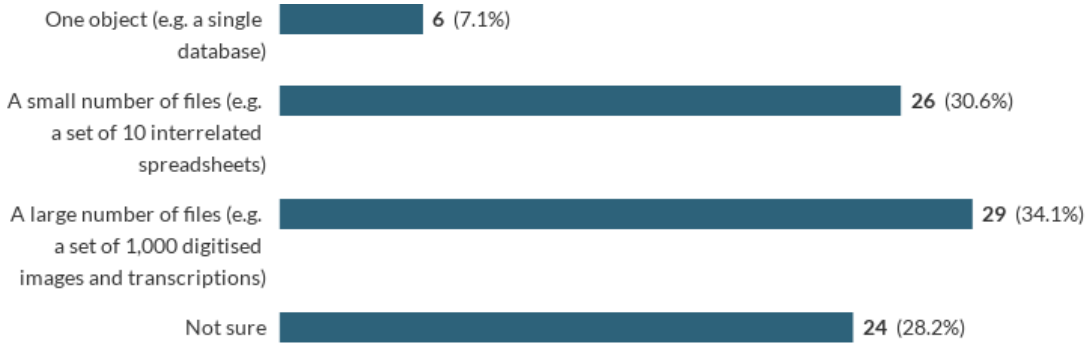


As we can see from the response above, the majority of researcher at Birkbeck are created datasets under 50GB. While 50GB is not a small amount of data in personal computing (in terms of photo storage, or music), this volume is comparatively manageable. As we go towards the 1TB and up to the 50TB+ levels, we require a different infrastructure to store and maintain this data.

We will more closely examine the data in a later section.

While volume is important for calculating costs, and pre-empting complications with things like archiving and sharing, the scale of digital objects is also of importance. A project with 1 single data set is much more easily to archive than a dataset of 1000 images. It means an increase in metadata, and can lead to issue around preserving file structure necessary to properly understand and reuse the data.

Q21a. How many digital objects did you create?



The survey shows that while there are smaller dataset currently being created (30.6%), the majority (34.1%) of dataset being created are of a larger scale. This presents us with a challenge as the current software being used to store and archive dataset, EPrints, is currently unable to represent non-flat folder structures.

Active Data Storage

Birkbeck IT Services provide network drives for general data storage. At the moment, due to the structure of the College and the individual school IT infrastructures, we cannot tell where data are currently being stored.

Locations for data storage are important when we look at best practice, backup, and planning ahead for the future.

Q22 Still thinking about your current or most recent project, where did you keep your data? (Please tick as many options as relevant.)

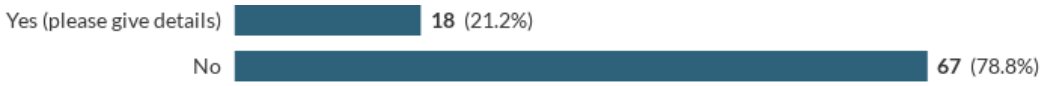
Storage Location		While working on them		To back them up		For long term storage		Total (%)
		count	% of total	count	% of total	count	% of total	
22.1	Local drive on my Birkbeck College computer	30	15.5	19	11.7	12	10.6	13
22.2	ITS proved personal Network drive	12	6.2	14	8.6	15	13.3	8.7
22.3	Local disk drive on a laptop/netbook	53	27.3	26	16	15	13.3	20
22.4	Portable storage device (e.g. external drive, USB disk)	26	13.4	35	21.5	22	19.5	17.7
22.5	Network servers dedicated to the project at Birkbeck College (NAS Box/ PC shared drive)	3	1.5	7	4.3	5	4.4	3.2
22.6	Network storage system/servers maintained by collaborating institution (NAS Box/ PC shared drive)	1	0.5	4	2.5	3	2.7	1.7
22.7	Shared storage area of the ITS servers (dedicated to the project team)	5	2.6	5	3.1	6	5.3	3.4
22.8	CD/DVD	1	0.5	2	1.2	7	6.2	2.1
22.9	Email system	18	9.3	9	5.5	4	3.5	6.6
22.10	Content/data management system operated by the project or School	1	0.5	1	0.6	2	1.8	0.9
22.11	Content/data management system operated by a project partner/collaborator	1	0.5	2	1.2	2	1.8	1.1
22.12	Web-based service, (e.g. Onedrive, Dropbox, Flickr, Google Drive, etc..)	36	18.6	33	20.2	18	15.9	18.5
22.13	Other	7	3.6	6	3.7	2	1.8	3.2

While there is a lot of information here, we can see the most popular locations highlighted in blue, with lighter blue for the second most popular locations.

The both most popular and second most popular are locations that would not be recommended while following best practices in RDM. USB storage can easily be lost, local storage is not backed up over the secure Birkbeck network, and Dropbox storage is not recommended unless approved by ITS.

ITS do have a OneDrive solution, however as we can see in the next section with a number of respondents paying for online storage with Dropbox, this service is not as well-known as it could be.

Q33 Have you ever paid for data storage



We can categorise and break down the paid as follows:

Q33a Please give further details of how much you paid for, and how it was costed.

Storage Location	Number Paying		Ave Amount Paid	Extrapolated College wide	
	count	% of total	(£)	count	(£)
Dropbox	11	78.6	85	75	6375
Hardware Storage	3	21.4	-	21	-
Zotero	2	14.3	37.5	14	525
Backblaze	1	7.1	60	7	420
Cloudme	1	7.1	60	7	420
Total staff paying for storage	14			124	7740

It has to be noted that we are working with a very small sample size in this area, so the extrapolation should not be seen as accurate; however it may give us some rough guidance. The results suggest that around 124 members of staff are paying for additional storage, at an overall cost of roughly £7740. This comes to £62 per staff member paying for storage.

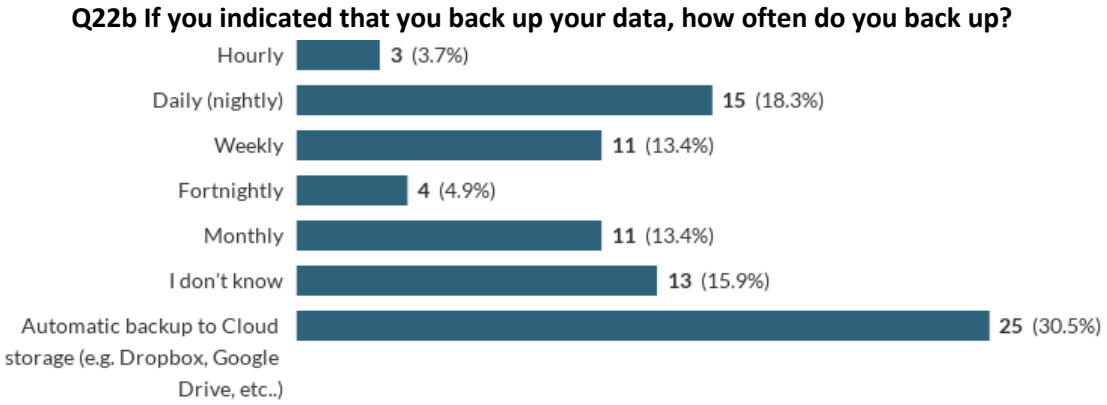
It should also be noted that this figure is per year, rather than outright. The outright hardware costs are not listed in the table, as the costs were in included in the answers given to the survey. Given the nature of the hardware reported (external hard drives, servers), we would expect that the costs here to be higher than the cloud style costs.

Finally, it should also be mentioned that at the moment we would not recommend using Dropbox for backup or long term storage. It has not proved to be secure, does not provide a backup guarantee, does not provide details of where your data is stored, and does not allow permanent deletion of sensitive files after a project has completed.

Backups

Best practices in RDM dictate that data should be store on secure spaces, encrypted where appropriate, and regularly backed up using the 3-2-1 Rule, as described by Peter Krogh in 2009[9].

As shown above, many of our researchers are choosing to back up on similar medium to their active data storage. This is of course not ideal, but the frequency of backup is important too.

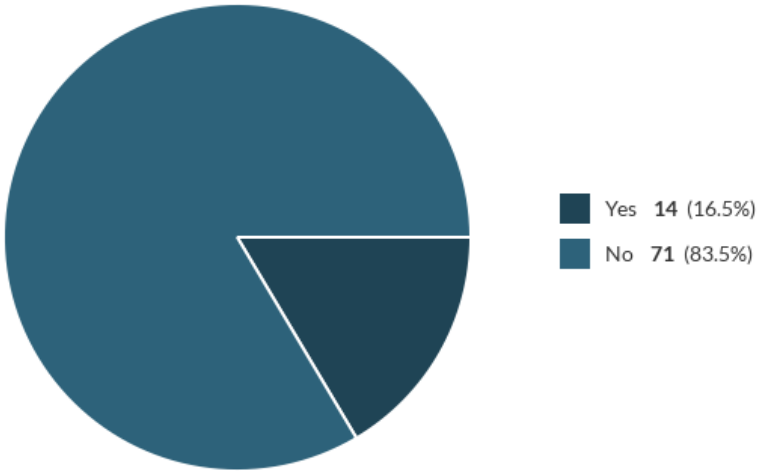


This shows that those who are backing up their data are doing it regularly. The 13.4% who back up monthly are at a higher risk of data loss however. Those who back up to the Cloud, are also taking a higher risk as there are often less guarantees that their data will be backup in alignment with good the good practiced as described by Peter Krogh.

[9] (Krogh, 2015)

Question 32 asks if the researcher has ever lost data.

Q32 Have you ever lost any Research Data?

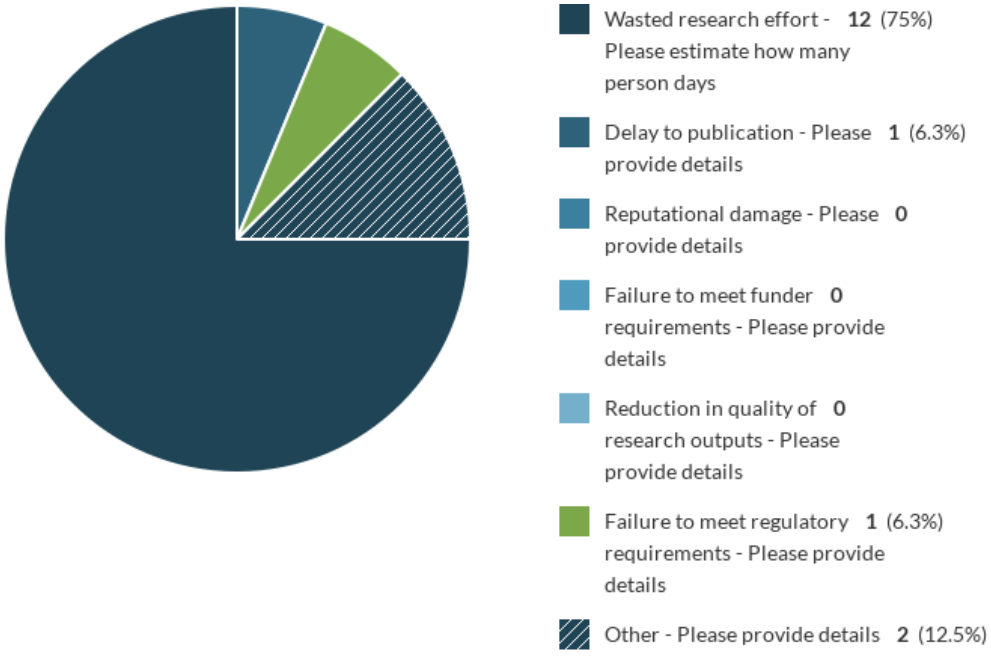


We then followed this up by asking what the cause and impact of the loss was. We can group the reasons for loss as follows:

Q32a What was the cause of the data loss? (e.g. hardware failure, human error, etc..)

Cause	Count
Hardware Failure	8
Human Error	5
Other	1

Q32b What was the impact of the loss? (tick all that apply)

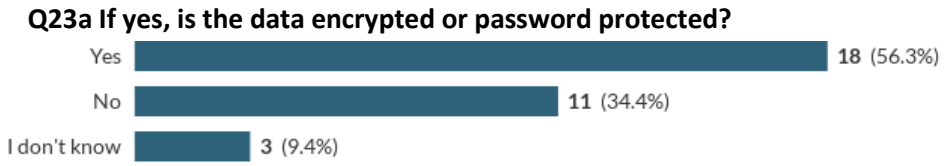


We did not get any estimates to how much effort was wasted, but that was the main impact of the data loss. Delays to publication and failure to meet regulatory requirements could be more important however.

Finally, by filtering on only those who have lost data, we can look back at Q22. Now we see 35.7% backup daily (over the 18.3% above). This may indicate that those who have previously lost research data back up more frequently.

Data Security

In questions 23 We asked “Does your Research Data contain any personally identifiable information or other sensitive data at any stage of the lifecycle? (prior to anonymization)”. 37.6% said yes, and went on to answer the follow up question:



Worryingly, we see in the responses here that only 56.3% of sensitive data stored at Birkbeck by researchers is encrypted or password protected.

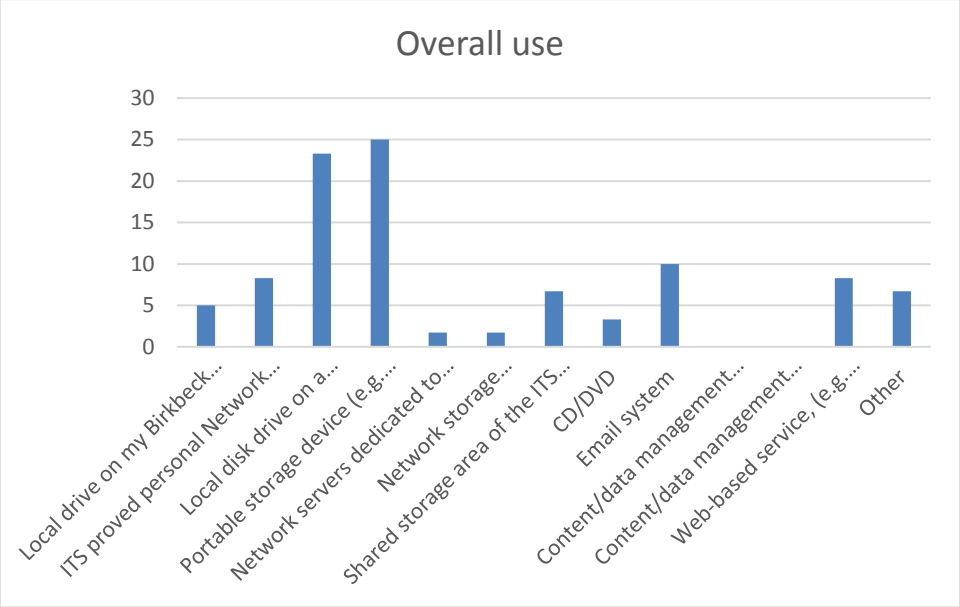
We followed up this question be asking those who said that they encrypted or password protected their data what software they used for this. This was a free text field, rather than a multiple choice.

We received 10 responses to this question, which we can categorise as follows:

Response type	**Count**	**Comment**
Adequately password protected in an encrypted drive or file system	1	The single instance of full encryption was with Carbon Copy Cloner, a proprietary backup solution. It is unclear how the encryption works if it is used here for active data.
Basic password protected at file level (Excel, Filemaker, etc..)	4	File level passwords are often suitable, but not as secure as using dedicated encryption software.
System level Password	1	System level passwords are easy to bypass, unless encryption is also used.
Cloud storage password	2	Cloud storage has been compromised in the past, and cannot be considered on the same level of security as dedicated encryption.
Physical data	2	Physical data that is kept in a locked safe may be secure from an disclosure viewpoint, but is not secure from loss through accidents like fire, as it cannot be backed up in the same way as digital data.

Sensitive Data

Q22 Still thinking about your current or most recent project, where did you keep your data? (filtered on Q23, sensitive and not encrypted)



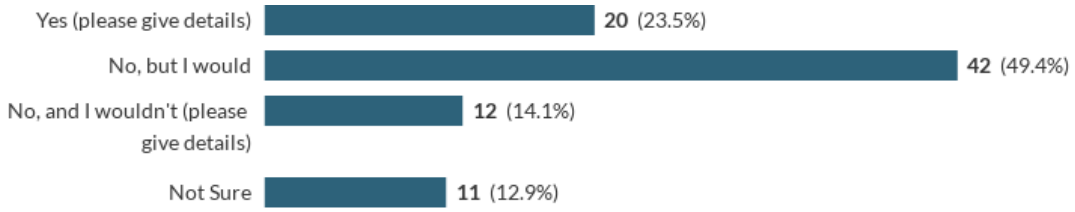
A total of 50% of all respondents who said they had sensitive data, but didn't encrypt it, store their data on either laptops or portable storage devices. This obviously increases the risk of the data being lost or stolen, and if the data is very sensitive it has numerous consequences.

1. It increases the risk to the participants of the research. If data is stolen if may be used to identify and target participants.
2. It increases the risk to the researcher. Losing participant data could damage their future career as a researcher.
3. It increases the risk to the institution:
    a. The College could suffer reputational damage if data was lost and then released publicly.
    b. When the General Data Protection Regulation (GDPR) comes into force, the responsibility for data security will lie with the College. Financial penalties can be applied for breaches.

Data reuse

We began by asking a broad question on if the respondent had ever reused data:

Q31. Have you reused someone else's data?



The response shows that around a quarter of our staff and PhD students have knowingly reused data (and rises to 31% if we only look at staff). Nearly half of respondents saying that they would reuse data.

Looking at the reasons given for reusing or not reusing, we see that most reuse is of public or already open datasets, while the reasons for not reusing data are more diverse in a smaller sample size.

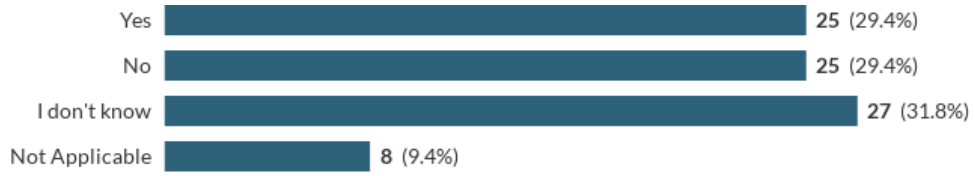
Data Archiving & Sharing

One of the reasons behind the rise of RDM, is the sharing of research data. With publicly funded data, this makes a lot of sense. Why should the public pay for data to be created twice, if it can be reused from a previous project.

There are of course many other reasons to share data. A paper from 2013 found that studies which make their data public receive on average 9% more citations[10]. Obviously not all data can or should be shared, but if all papers which could share their data did, a 9% increase in citation would not be an insignificant amount.

To look at our academics' attitudes toward sharing, we asked about their intention to deposit data at the end of their project.

Q25. Do you intend to deposit your data at the end of your current project?



We can see from the above chart that there is a split between the "yes" and "no" responses. The largest however, are the "I don't know". This suggests that with more training and advice, we may see more enthusiasm for depositing data.

[10] (Piwowar and Vision, 2013)

We also asked where would they actually put their data at the end of their project:

Q25a If yes, where do you intend to deposit your data?

Locations	Count
Subject Repository	6
BIROn	4
Don't Know	3
BiRD	3
Local Storage	2
Commercial Repository	2
ITS server	2
Orbit	1
Open Repository	0

As can been seen from the above list, there are a lot of locations where our academics intend to deposit data are inappropriate for long term storage.

The only locations listed above that we would recommend for long term storage of data after a project has been completed are BiRD, Open Repositories, and Subject Repositories.

The high prominence of BIROn suggest that we may need some signposting within BIROn to direct depositors to the correct repository.

Of the 22.4% who have previously deposited data, the most common location reported was BIROn. Again, if we have sign posting within the repository we may see a higher rate of deposit in BiRD.

Q28a If so, where did you deposit the data?

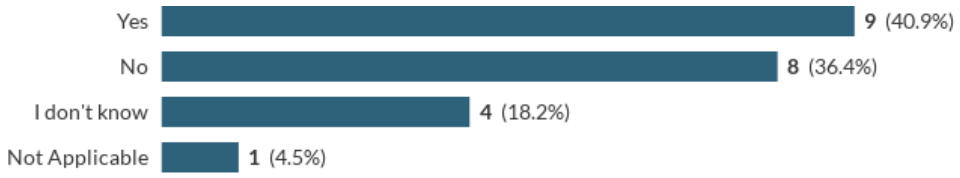
Response	Count
BIROn	5
ESRC/ReShare/UKDA	4
Subject Archive	4
Journal Archive	1
Figshare	1
BiRD	1

Data Archiving & Sharing – Funder Compliance

A large portion of the respondents stated that they are in receipt of funding from one of the many bodies who now have policies stating that data should be deposited and shared at the end of a project.

We then limited the responses to just those funders who expect data to be made available after the competition of the project and looked at the deposit intentions:

Q25. Do you intend to deposit your data at the end of your current project?
(Funder expected deposit only)



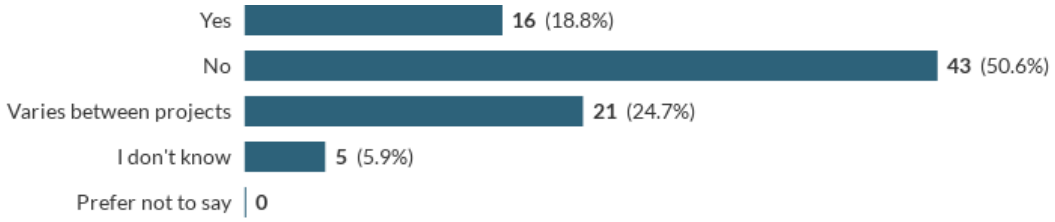
Here we can see that, while the funders expect data to be deposited, only 40.9% report that they intend to do so.

Looking at the reasons given for not depositing, there is a mixture of general reluctance, anonymity concerns, and ownership issues. The first should not be an issue in depositing data when the funder of the researcher is requesting it, and the second two should be address in the DMP.

Active Data Sharing

Whether due to distant collaborations, or just due to the nature of modern researcher, many projects require the sharing of data during the active phase of the project.

Q35. Do you share the Research Data that you create/manage beyond the project team during the life of a project?



We are also interested in how this active data is shared, and who with.

Q35a. If yes, who do you share your data with during the project life?

Party shared with	Count
One or more members of the research group	24
Collaborating partners at other institutions	24
Anyone who expresses an interest	10
One or more members of the department	6
Publisher(s)	6
Wider public	5
Funder(s)	4
Anyone within the School	3
Third party Data Provider / Data Creator	1
Other	1

Q35b. If yes, how do you share your Research Data?

Type	Count
Cloud*	24
Publish online	19
Email	18
Physical	17
Private Server	9
Code sharing platform	4
Other	4
Institutional File Share	2

*Dropbox like services

We can see that the most common parties active data is share with are other academics, either within the project team locally, or at collaborating institutions. We can see that the most popular method of sharing active data is cloud services such as Dropbox. This is not a great surprise considering how easy to use these platforms are.

Finally, we asked what common applications were used by our researchers. The top 10 were as follows:

Application	Count
MS Excel	40
SPSS	22
MS Word	16
Matlab	9
Custom Applications	6
R	6
Not Applicable	5
Stata	4
NVivo	3
PDF Maker	3

Training Requirements

As part of The Colleges commitment to complying with EPSRC requirements for data management[11], we plan to run RDM workshops to provide training to staff and students. While some of these workshops could be suitable for all, some more tailored versions could be better suited to disciple or career stage.

We included some questions around training requirements in the survey to allow staff to tell us where they think they might need training, and what training they have already received.

Q12. Have you received any training or support on Research Data Management?

The yes responses then went on to describe the training they had received, which was either not a workshop or training event in the sense we are look at, or were at previous institutions.

Q13 Would staff development or training be useful to you in any of the following areas relating to Research Data Management?

Training Description	Very interested	Might be interested	Not interested	No opinion
An introduction to Research Data Management	21	40	19	3
Citing Research Data	7	35	36	4
Collaboration and sharing of data	14	37	26	5
Copyright and intellectual property rights within a data context	18	37	21	7
Data anonymisation	17	25	36	4
Data licensing	10	22	37	11
Developing Data Management Plans (DMPs), Technical Plans, and Data Sharing Plans	16	30	27	9
Documenting your Data	17	30	25	10
Ethics, consent and legal issues with Research Data	16	31	29	6
Funder requirements for Research Data Management	14	37	24	7
Guidance on costing Data Management in grant applications	24	28	24	5
Long-term storage of your data	27	36	16	2
Publishing Research Data	15	38	24	2
Security of data	16	35	25	4
Sharing your Research Data	17	42	17	4
Software carpentry	9	18	32	21
Support in data selection, metadata creation and licensing for preservation	15	28	28	10

[11] https://www.epsrc.ac.uk/about/standards/researchdata/expectations/

By looking at the scores we can see the following are the top responses:

Top 5
Long-term storage of your data
An introduction to Research Data Management
Sharing your Research Data
Copyright and intellectual property rights within a data context
Guidance on costing Data Management in grant applications

Based on this we can look at developing a training workshop programme.

Data (current, future, and at risk)

One of the most important parts of the survey was to try and understand the existing level of data, it's nature, and current state help by the college. By knowing this we can better provision for future data, and mitigate the risk of losing the data that we already have.

Volume and Nature

We asked our participants to tell us how much data they are storing, and to estimate how many files they have created. The volume is important for costing storage needs going forwards, and the scale is important for tailoring archival practices.

When we combine the answers for Q21 and Q27 (These questions ask about current data volumes and data volume from past projects still stored), we can get an idea of the total data stored at The College, and who is storing the most data.

There are a number of issues with approach. Firstly, the main data producers may be more likely to respond to a survey on research data. This would skew the data volume upwards. Secondly, number of respondents across the departments is not equal. We are looking here at an average overall, not per department.

Given these issues, the following figure can only be thought of as a rough estimate, and not an accurate prediction of data volume. This is important as Science appeared to create far more data than Law or Social Sciences, History, for example.

The overall volume of data stored by The College may be as high as of 3.52 Petabytes (See Appendix C).

Even if we exclude the 5 largest data creators/holders, (those with the response >50TB), we still see that we could currently have around 1.66PT of data stored.

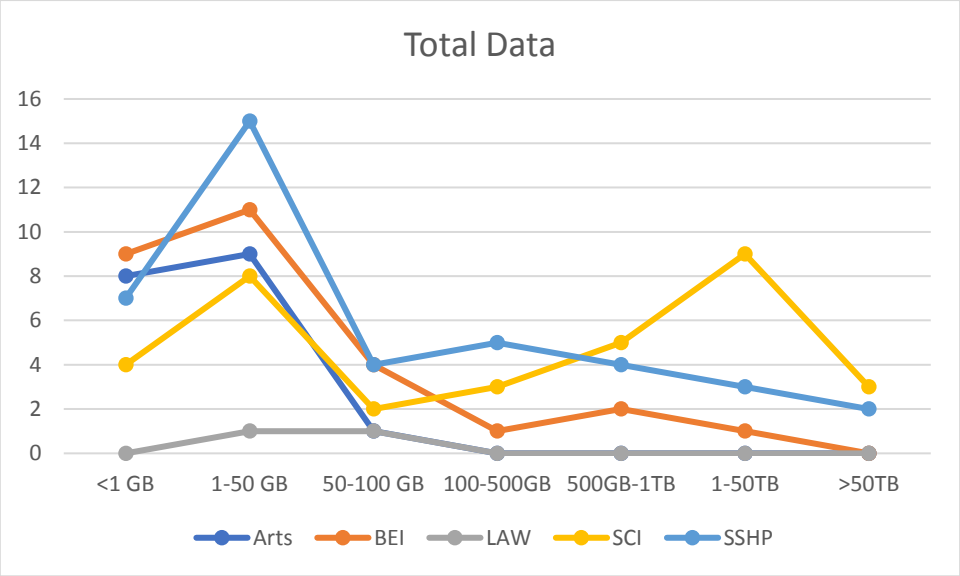
So, to summarise:

Totals	**Extrapolated Volume Stored (Active & Historic)**
Excluding largest creators (over 50 terabytes)	1.66 PB
Including largest creators	3.52 PB

Volume by School

We can however break down the results further, to find which schools are the primary data creators and holders within The College.

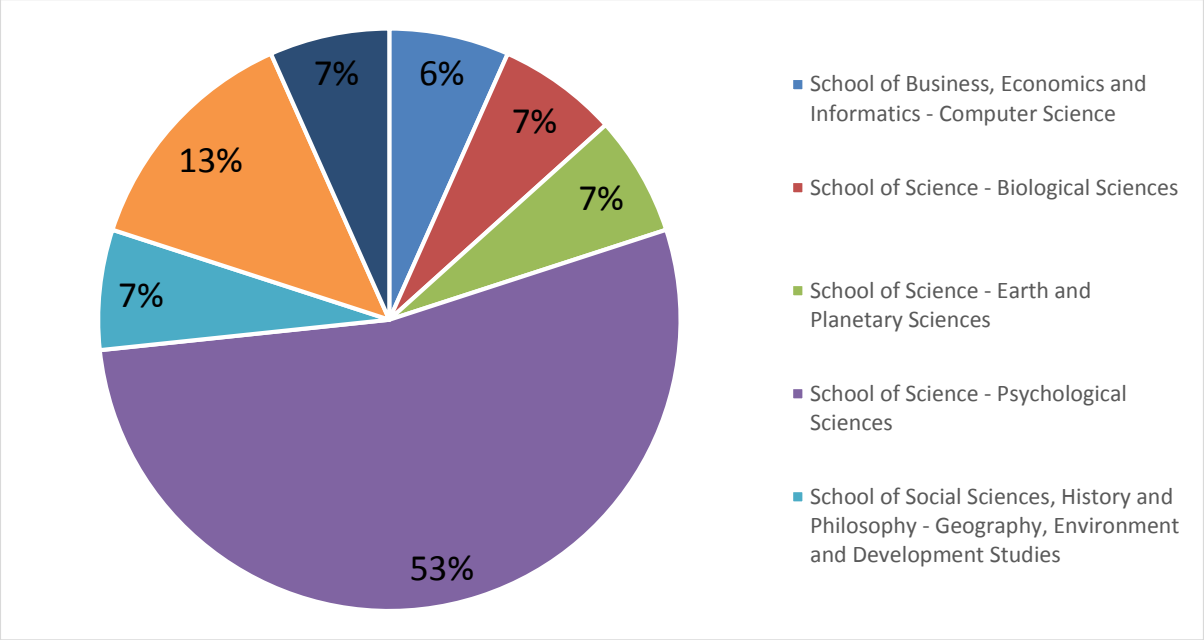
Q21 & Q27 combined with Q1



We can clearly see (based on the responses received) that The School of Science is our largest data creator and holder, followed by The School of Social Sciences, History and Philosophy. In fact, between them they account for around 96% of data created by The College. However, please note the low response rate from the School of Law so these figures are less reliable.

Looking at our largest data producers (those who responded that they create or store data above 1TB), we can see that some departments/schools create far more than others:

Q1 Which School and Department are you located in? (limited by responses to Q21 & Q27, data volumes over 1TB)

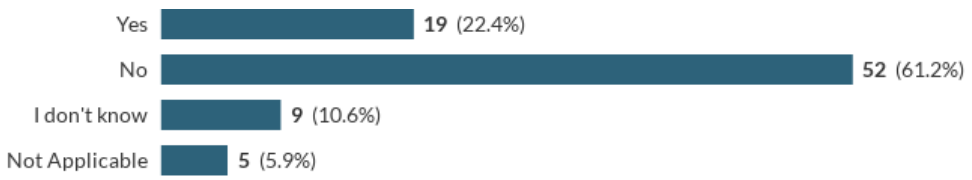


It is no surprise that the producer/holder of the largest data at Birkbeck are the School of Science, with 66.7% of all data over 1TB.

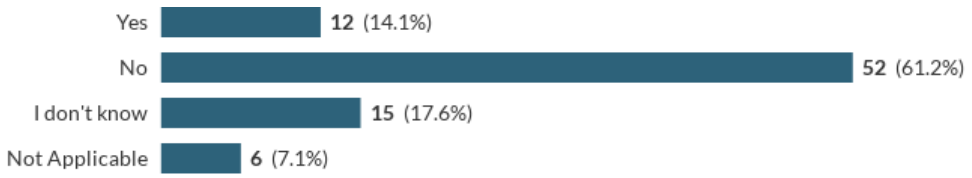
Existing Archival Data

We asked a series of questions looking to establish what volume of data our researchers thought they might like to archive in the future. This may help us to decide whether we should be aiming to archive this data, and if so what space we would need.

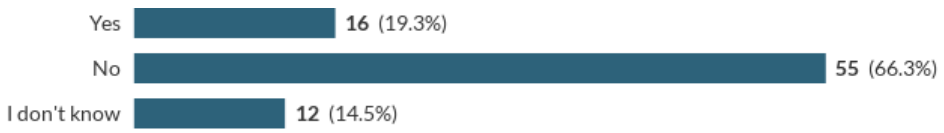
Q28. Have you deposited/archived any of your previous research in a data repository?



Q29. Do you have any Research Data from a previous project you would like to deposit/archive?



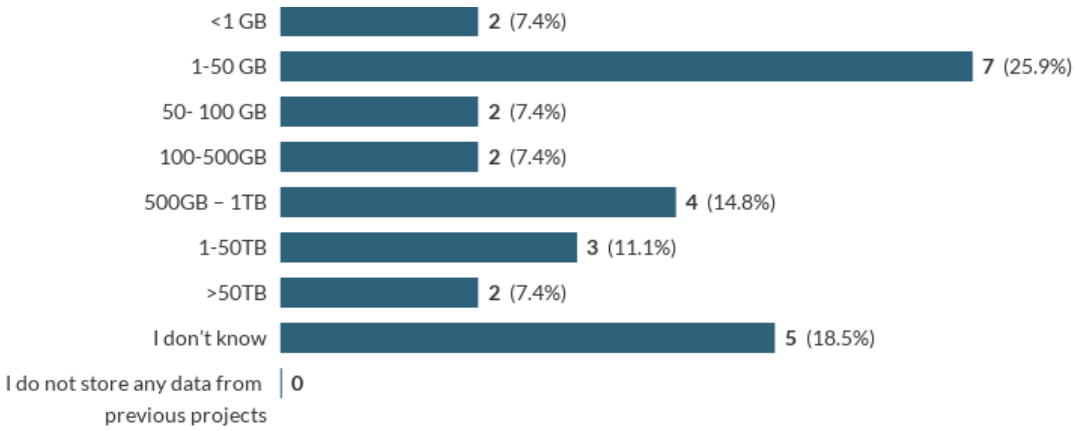
Q37. Do you have any other datasets (historic or not supporting your current research) that you may like to place in a secure repository, hosted by the College Library?



From the above we see that the majority do not feel like they have any data they would like, or should deposit in an archive, Birkbeck's or otherwise.

Limiting the responses to those who already have deposited in the past (Q28), have data they'd like to deposit (Q29), or are interested in depositing in BiRD (Q37), we see the following data volumes:

Q27 Thinking about your past or previously conducted research, estimate what volume of data you are still storing. (see "more info" for guidance)



(this was also limited to Staff, as current PhD students would not be eligible to deposit their historic data)

Excluding the very large data, we can extrapolate these results to find that we may have as much as 385.65TB of research data from previous projects that our researchers would want to archive.

It is very hard to judge the accuracy of this, due to the very small sample size and the variance across disciplines.

Comparisons

Due to the deliberate similarities with other surveys conducted at other HEIs, we can make some comparisons to our own.

We will mainly be drawing comparisons with:

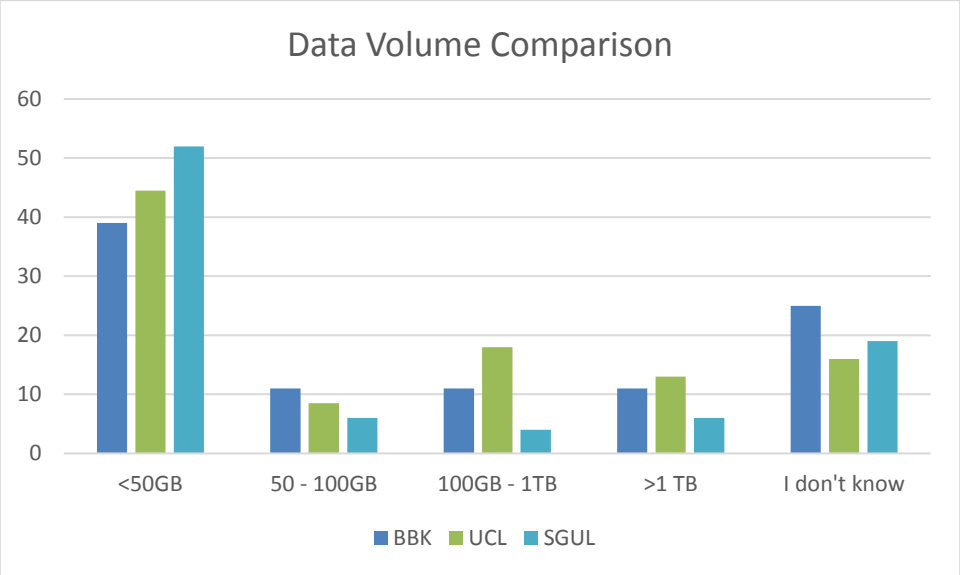
- The UCL Research Data Management Survey 2016[12]
- Jisc Data Asset Framework Surveys 2016[13]
  This survey combines results from set DAF conducted at:
  - CREST
  - Lancaster University
  - Plymouth University
  - The Royal College of Music (RCM)
  - The University of Cambridge
  - The University of St Andrews
- Research Data Management at London School of Hygiene and Tropical Medicine: Web Survey Report 2013[14]
- Research Data Management at St George's, University of London Web survey analysis 2015[15]
- The 2014 DAF Survey at the University of Sheffield[16]

In the following four sectors we compare our Birkbeck results to the above, where the questions are similar enough to allow us to do so.

[12] (Fellous-sigrist, 2016)
[13] (Open Exeter Project Team, 2012)
[14] (Knight, 2013)
[15] (Basford, 2016)
[16] (Cox and Williamson, 2015)

Data Volumes

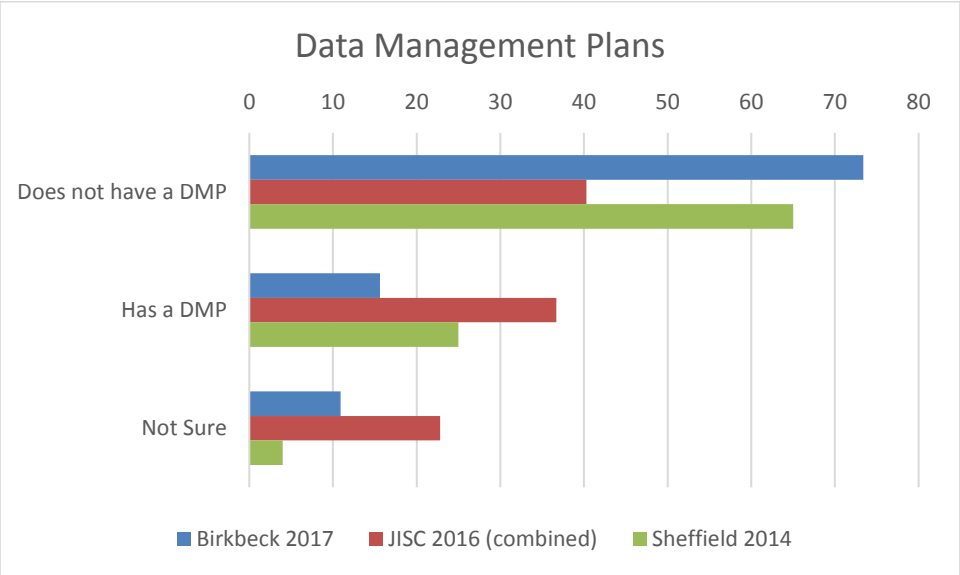
Numerous surveys have asked about the volume of data being generated or stored.



From the chart above we can see that Birkbeck is quite similar to UCL and SGUL. ULC may reflect our research interests more closely, while SGUL is more similar in size. It's reassuring that we are not generating more data per researcher than other similar institutions.

Data Management Plans

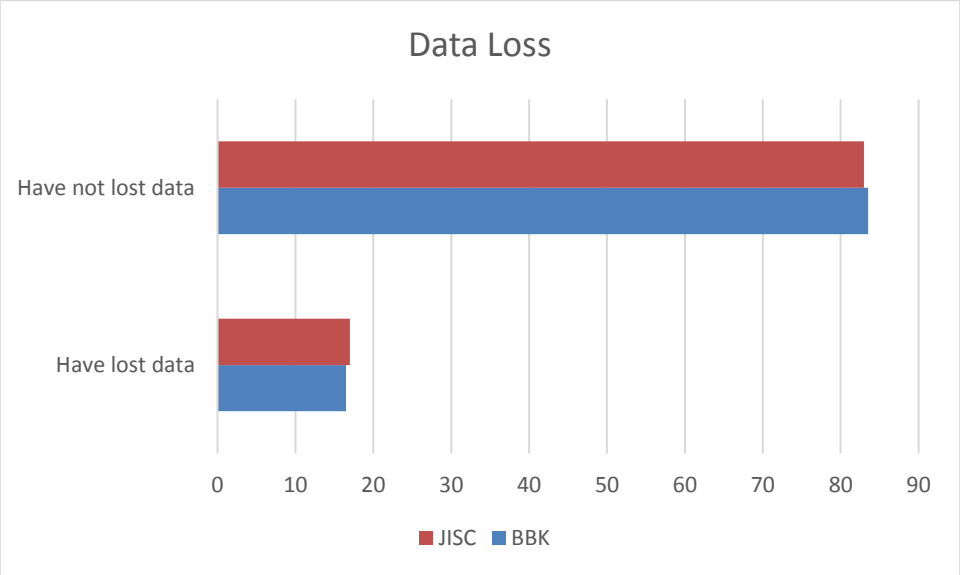
As show in the Policy Awareness section, researchers are Birkbeck often do not have Data Management Plans for their research. In this comparisons section, we wanted to find if this is common across the sector. To achieve this, we took the results of our survey and compared them to the results of similar questions asked in the Jisc 2016 study of multiple institutions, and the 2014 Sheffield survey.



As can be seen above, the 2016 Jisc survey has the most positive results. Looking slightly further back in time, Sheffield managed similar responses to Birkbeck. However, it is clear that we are slightly behind the institutions Jisc surveyed in this area, and this must be addressed.

Data Loss

For our survey, we used the exact same question as Jisc, and got the following results:

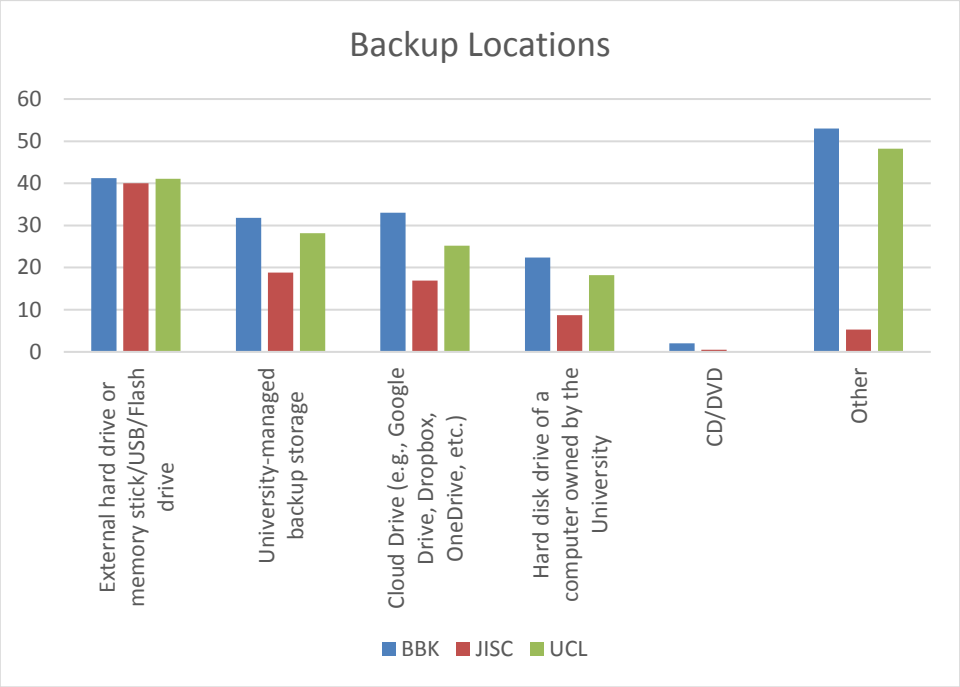


Our results very closely match the combined Jisc outcome, suggesting that 16.5 - 17% of academics losing data is fairly standard across the sector.

Our main reasons for loss were “Hardware failure”, and “Human error”. This matches the results Jisc had.

The main impact of the loss was “Wasted research effort”, which is the same for the Jisc survey. They reported higher numbers in “Delay to publication” and “Reduction in quality of research outputs”. This suggests that we should be aware of these are potentially important, given we had a much smaller sample size.

Backup Locations



From the above we can see that Birkbeck researchers are more reliant on Cloud storage for backup than others in the sector. There is some issue with the above in that Jisc have a very small “Other” level, which suggests that the way the questions was framed may have been different. This could be due to the way the survey had to be institution neutral.

Conclusions

1. How have attitudes towards Open Access changed over the years since the previous study in 2011?

We see that there hasn't been much move towards acceptance of Open Access. Researchers are less enthusiastic about OA and OA journals now that they are more commonplace.

The free text responses we got to the point to concerns over costs, the hybrid nature of some journals, and the integrity issues paying to publish create.

Awareness of the institutional repository BIROn has improved, so current efforts to promote have been very successful and should be continued.

2. What are the current attitudes towards Open Data and RDM?

Three quarters of respondents believe data should be shared freely where possibly. This is very positive. While staff awareness of policy requirements, of existing infrastructure are low, and most staff have not revived training in RDM practices, their enthusiasm should make improving these numbers possible.

3. How aware are Birkbeck researchers regarding internal and external RDM policy?

As shown in the Policy Awareness section, our researchers are not aware of the majority of policies that affect their research, including local Birkbeck policies.

We should aim to improve DMP & data deposit rates for all research, which would help us comply with these policies.

4. What are Birkbeck researchers current RDM practices? (This includes details on volume, security, location, accessibility, etc..)

Our researchers are storing a lot of data in places we wouldn't recommend. But a lot of it is being backed up.

The responses also show that up to 30% of data created at The College is not in a condition to easily be shared. With storage space on campus at a premium, an argument could be made for a push toward digitisation and disposal of large historic paper based datasets, whilst recognising that there are also issues with this approach. Many types of physical data can never be digitised, but a catalogue record of these would make sharing and sustaining these resources easier in the long term.

Our researchers also produce datasets with a large numbers of files. This is difficult to represent accurately in our current repository software EPrints. Going forward this should be a consideration when we decide on continued development of the system, and other options on the market.

5. What are the training needs that our researchers think they have?

The training needs are clearly laid out in the section above. Based on their responses we can plan for running workshops in the near future.

6. How does Birkbeck compare with other institutions in this area?

We seem to be in roughly the same place as the other institutions who had recently been surveyed. Our reliance on the cloud as a solution to storing, backing up, and sharing data is higher when compared to some, and this is a significant concern.

Recommendations

Kirsten Briney describes data management as involving, planning, documenting, organising, analysis, securing, storing and backing-up during a project and after a project, sharing, and finding and reusing data.[17] Supporting at all stages should be the target for a research data management support service, breaking this down to pre-project support, active project support, and post-project support, all as business as usual activities.

We can use the responses from the survey, and the analysis in the Outcomes section to recommend some actions for the service going forwards:

Pre-Project Support

1. We should continue to suppose researchers request for help with data management plans, one-to-on or though workshops. This will help our compliance and improve out communications without researchers.
2. We should develop a template DMP in the DMPonline tool and publicise its existence.
3. We should aim to provide sample plans, or access to successful plans completed by other Birkbeck researchers.
   Identifying sensitive data is part of the DMP process. Given the results we presented in the Sensitive Data, with 50% of respondents who have sensitive data not storing it correctly, it makes sense to plan for this with a DMP. We should consider our ethical approvals processes in the context of data management plans, with particular consideration of the approval of projects which are classified as routine and therefore signed off by the researcher without further scrutiny.

Active Project Support

Our researchers are not currently storing tor backing-up their data in the most suitable places. This can be seen in more than one section of the Outcomes.

With 17% reporting data loss at some stage (which has been shown and reported to cause delays to publication, wasted research effort, and lower quality outputs[18]), we need ensure our academics have infrastructure in place to support them, and importantly that they are aware of the tools provided by ITS in this area.

1. We should work with ITS to provide suitable storage for data sharing, wither this be OneDrive or SharePoint.
2. We should promote the facilities provided by ITS, and the benefits of using managed back-up solutions.
3. We should provide information on how to prepare your data for storage during the project. This especially applies to those projects where funders would like the data to be deposited

[17] (Briney, 2015)
[18] (Johnson *et al.*, 2016)

within a certain timescale, relating to when the data was completed, rather than when the project was completed.

These all imply a more complete website in the library, and working more collaboratively with ITS.

Post-Project Support

In some cases, lack of knowledge of funder policy could impact on future funder opportunities, if the correct post-project procedure is not followed.

1. We should consider contacting PIs whose projects that have recently been completed, to offer assistance with depositing their data.
2. We should contact self-depositors to the publications repository BIROn, so they are aware they can link their data to their publications.
3. We should also add signposting to BIROn to ensure that self-depositors are aware of BiRD.

Business as Usual

To comply with funder requirements, we should run regular training sessions.

1. We should run at least 3 workshops per session. The most popular replies, and the area where most knowledge is needed, which easily translate into workshops are the following:
    - Intro to RDM (Introduction, sharing data, backups, securing, DMPs, etc..)
    - Storing and Sharing data (How to use BiRD, and other data repositories)
    - Data Management Plans and how to use DMPonline
2. There were also suggestions that do not fit quiet so easily into workshops run by the library. These could be run in collaboration with Research Grants and Contracts, or the Legal and Governance office.
    - Copyright and IP
    - Costing

We should try and engage more with our academics to raise general RDM awareness, along with awareness of the service and the policies.

1. Attend departmental meetings with subject librarians where possible.
2. Run a regular “Drop In”. These are not well attended elsewhere, however so long as they do not impact the running of the service they may still be useful.
3. Develop promotional materials for the service to aid the improving of awareness. This may include flyers, posters, and larger pull out poster boards for workshops/drop-ins.
4. Continue to run promotional events such as Bloomsbury Data Week, to raise awareness of RDM.

Lessons Learned

While the survey was successful overall, and returned useful results for us to improve the services we offer, there were some lessons learned for similar questionnaires in future.

1. Avoid formatting questions using the Bristol Online Survey “Scale” option if you are not familiar with sorting through unformatted data.
2. Some questions were of limited use to the PhD students who responded to the survey. Some “routing” should have been added to avoid this.

Data Access Statement

Anonymised data is available on the Birkbeck Data Repository (BiRD) here: https://doi.org/10.18743/DATA.00012

Acknowledgements

The Survey was based on work by Data Asset Framework (DAF), which was developed by HATII at the University of Glasgow with help from the Digital Curation Centre, and funded by Jisc.

Further questions were added to the basic DAF, with inspiration coming from UCL, UEL, LSHTM, SOAS, and SGUL. Support from within Birkbeck was also essential in tailoring the questions and structure to our requirements.

References

Basford, J. (2016) 'Research Data Management at SGUL Web survey analysis : Sept – Nov 2015', (January), p. 27. Available at: researchdata@sgul.ac.uk.

Briney, K. (2015) *Data management for researchers : Organize, Maintain and Share your Data for Research Success*. Pelagic Publishing, UK.

Cox, A. and Williamson, L. (2015) 'The 2014 DAF Survey at the University of Sheffield'.

Fellous-sigrist, M. (2016) *UCL researchers and their research data: practices, challenges & recommendations Report on the 2016 RDM Survey*, *UCL Library Services: London, UK.* UCL Library Services. Available at: http://discovery.ucl.ac.uk/1540140/ (Accessed: 31 March 2017).

Johnson, R., Chiarelli, A. and Parsons, T. (2016) 'Data asset framework (DAF) survey guidance 2016'. doi: https://doi.org/10.6084/m9.figshare.3796305.v4.

Johnson, R., Parsons, T., Chiarelli, A. and Kaye, J. (2016) *Jisc Research Data Assessment Support - Findings Of The 2016 Data Assessment Framework (Daf) Surveys*. doi: 10.5281/ZENODO.177856.

Jones, S. (2011) *How to develop a data management and sharing plan*, *DCC How-to Guides*. Edinburgh. doi: http://www.dcc.ac.uk/resources/how-guides.

Knight, G. (2013) *Research data management at LSHTM: web survey report*. Available at: http://blogs.lshtm.ac.uk/rdmss/files/2013/04/LSHTM-RDM-Web-Survey-Report.pdf.

Krogh, P. (2015) *Backup Overview*, *American Society of Media Photographers*. Available at: http://www.dpbestflow.org/backup/backup-overview (Accessed: 10 April 2017).

Open Exeter Project Team (2012) 'Summary Findings of the Open Exeter Data Asset Framework Survey'. Available at: http://hdl.handle.net/10036/3689.

Piwowar, H. A. and Vision, T. J. (2013) 'Data reuse and the open data citation advantage', *PeerJ*. PeerJ Inc., 1, p. e175. doi: 10.7717/peerj.175.

Appendix A

No.	Question
1	Which School and Department are you located in?
2	Are you part of any Birkbeck Research Centre or Institution?
2a	Are there any errors in the above list of Research Centres and Institutions?
3	Which of the following best describes your research experience
3a	If you selected Other, please specify
4I	Open Access: How do you feel about the principles of Open Access?
4II	Open Access: How do you feel about using Open Access repositories?
4III	Open Access: How do you feel about publishing in Open Access journals?
4a	Open Access: Do you have any further comments to add about the principles of Open Access?
5	Do you feel you understand the Open Access requirements for inclusion in the next REF?
5a	If No or Not Sure, how can we better communicate these requirements?
6	Do you feel you understand the differences between Gold and Green Open Access?
7	Are you aware of The College repository, BIROn?
7a	Do you currently make any of your publications available in BIROn?
7a.i	Why are you not making your publications available in BIROn?
8	Do you deposit your own publications? (self-archive)
9	Who do you think “should” own the copyright of research publications?
9a	If you selected Other, please specify:
10	Who do you think “should” own the copyright of Research Data?
10a	If you selected Other, please specify:
11	Do you think Research Data should be shared openly where possible? (Data that is not commercial sensitive, or impossible to anonymise)
11a	If not, could you explain why not?
12	Have you received any training or support on Research Data Management?
12a	If yes, please describe the training you received
13	Would staff development or training be useful to you in any of the following areas relating to Research Data Management?
13a	Please describe any other type of RDM related training you think might be useful for staff or research students
14	Have you ever create a Data Management Plan, Technical Plan, or Data Sharing Plan for any research you have undertaken here at Birkbeck?
14a	What are your reasons for not having created a Data Management Plan?
14b	What are your reasons for having created a Data Management Plan?
15	If you have previously completed a DMP/Technical Plan/Data Sharing Plan, as part of an application for funding, do you think your funder seriously considered your responses before reaching a decision regarding the bid?
16	Are you the Principal Investigator (PI) on your current or recent project?
17	How was this project funded?
18	Do you have a data management plan (DMP) for your current research?

19	Describing your Research Data
19a	What data 'types' are involved with your research? (please think more generally if your current project does not represent your regular research interests)
20	Do you store any non-digital Research Data?
20a	What kinds of non-digital Research Data do you store?
20b	Do you create digital copies of this data?
21	Estimate what volume of data are you creating?
21a	How many digital objects did you create?
22	Still thinking about your current or most recent project, where did you keep your data?
22b	If you indicated that you back up your data, how often do you back up?
23	Does your Research Data contain any personally identifiable information or other sensitive data at any stage of the lifecycle? (prior to anonymization)
23a	If yes, is the data encrypted or password protected?
24	What, if any, legislation policies or other rules influence how your Research Data is stored, managed and/or shared?
25	Do you intend to deposit your data at the end of your current project?
26	In past or previous research, where did you keep your data? Please tick as many options as relevant.
27	Thinking about your past or previously conducted research, estimate what volume of data you are still storing.
28	Have you deposited/archived any of your previous research in a data repository?
29	Do you have any Research Data from a previous project you would like to deposit/archive?
30	Who has the Intellectual Property Right for your research data?
31	Have you reused someone else's data?
32	Have you ever lost any Research Data?
32a	What was the cause of the data loss?
32b	What was the impact of the loss?
33	Have you ever paid for data storage?
34	Is any of your data (past or current) part of a collaboration?
35	Do you share the Research Data that you create/manage beyond the project team during the life of a project?
35a	If yes, who do you share your data with during the project life?
35b	If yes, how do you share your Research Data?
36	Please list the three most common tools you use to create and/or manipulate your Research Data (e.g. Stata. MySQL, MathWorks, Microsoft Excel).
37	Do you have any other datasets (historic or not supporting your current research) that you may like to place in a secure repository, hosted by the College Library?
38	Do you have any other comments, views, experiences, or advice on Research Data Management at Birkbeck that you would like to share in order to improve RDM support in the College?

Appendix B.1

Question 4 (Staff)	2011	2017	Change	2011	2017	Change
	Agree or Strongly Agree			Disagree or Strongly Disagree		
How do you feel about the principles of Open Access?	89.4	81.3	-8.1	3	6.3	3.3
How do you feel about using Open Access repositories? (Repositories which make versions of articles freely available)	83.3	87.5	4.2	3	4.7	1.7
How do you feel about publishing in Open Access journals? (Journals which do no restrict access to articles)	63.1	68.7	5.6	15.4	7.8	-7.6

Appendix B.2

Question 7 & 8 (Staff)	2011	2017	Change	2011	2017	Change
	Yes			No/Not Sure		
Are you aware of The College repository, BIROn?	43.9	96.9	53	56.1	3.1	-53
Do you currently make any of your publications available in BIROn?	46.7	91.9	45.2	53.3	8	-45.3
Do you deposit your own publications? (self-archive)	12.5	79.7	67.2	87.5	20.3	-67.2

Appendix C

Overall data stored at The College. This combines Q21 & Q27 (active and historic research data). The totals should be considered very rough as we do not know if those who are more likely to respond to a survey such as this are also more likely to hold large volumes of data.

The second total attempt to avoid this issue by excluding the very largest data creators and holders.

	responses	**at with 438**	**rough vol**	**est data volume for college**
<1 GB	28	144	0.5	72
1-50 GB	44	227	25	5675
50-100 GB	12	62	75	4650
100-500GB	9	46	250	11500
500GB-1TB	11	57	750	42750
1-50TB	13	67	25000	1675000
>50TB	5	26	75000	1950000
I don't know	39	201		
I do not store any data from previous project	12	62		
total	173	891		3689647
		Totals	GB	3689647
			TB	3603.17
			PT	3.52
		Totals	GB	1739647
		(Excluding >50TB)	TB	1698.87
			PT	1.66